SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Computational life sciences
• RESEARCH PAPER •

November 2014 Vol.57 No.11: 1080–1089
doi: 10.1007/s11427-014-4744-9

A generative model of identifying informative proteins from dynamic PPI networks

ZHANG Yuan[1], CHENG Yue[1], JIA KeBin[1*] & ZHANG AiDong[2]

[1]*Department of Electrical Information and Control Engineering, Beijing University of Technology, Beijing 100124, China;*
[2]*Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260-2500, USA*

Received May 23, 2014; accepted August 1, 2014; published online October 17, 2014

Informative proteins are the proteins that play critical functional roles inside cells. They are the fundamental knowledge of translating bioinformatics into clinical practices. Many methods of identifying informative biomarkers have been developed which are heuristic and arbitrary, without considering the dynamics characteristics of biological processes. In this paper, we present a generative model of identifying the informative proteins by systematically analyzing the topological variety of dynamic protein-protein interaction networks (PPINs). In this model, the common representation of multiple PPINs is learned using a deep feature generation model, based on which the original PPINs are rebuilt and the reconstruction errors are analyzed to locate the informative proteins. Experiments were implemented on data of yeast cell cycles and different prostate cancer stages. We analyze the effectiveness of reconstruction by comparing different methods, and the ranking results of informative proteins were also compared with the results from the baseline methods. Our method is able to reveal the critical members in the dynamic progresses which can be further studied to testify the possibilities for biomarker research.

dynamic protein-protein interaction network, abnormal detection, multi-view data, deep belief network

Citation: Zhang Y, Cheng Y, Jia KB, Zhang AD. A generative model of identifying informative proteins from dynamic PPI networks. Sci China Life Sci, 2014, 57: 1080–1089, doi: 10.1007/s11427-014-4744-9

The emerging of translational bioinformatics (TBI) builds a bridge between fundamental findings of bioinformatics and clinical practices [1]. One of the most important subjects of TBI is to identify the informative genes or proteins related to disease development and biological processes [2,3] which can provide important evidence for biomarker detection in specific biological processes.

A biological process is a complexity of spatial and temporal interactions among innumerable molecules. Dynamic biological network mining has attracted increasing attention from biologists in the past few years [4–6] because they capture the temporal relationships between proteins. The informative proteins inside dynamic protein-protein interaction networks (PPINs) are defined as the proteins with dramatic topological changes during the biological processes. This assumption is based on the definition proposed by Han et al. [7], in whose study, dynamic patterns of protein interactions are discovered and proteins are divided into two categories, that is, the highly positive co-expressed proteins which tend to form the most static modules appearing at all time and the hubs at the center of which being referred to as "party" hubs; and the less positive coexpressed proteins interactions appearing at particular time points, inside which the hubs therefore being referred to as "date" hubs that are believed to cause dynamic interactions and induce possible aberrant pathways and molecular disorders. Taylor et al. [8] also observed multi-modal distribution of correlation coefficients of gene expression using curated sources from literature. Among the "party" and "date" hubs, the latter ones are more essential to global connectivity and functions that

*Corresponding author (email: kebnj@bjut.edu.cn)

cells possess which are also the possible informative proteins that we want to discover.

The traditional methods of informative protein detection mostly focus on the topological characteristics of PPINs [9–11], and the topology-based metrics, for example, degree and clustering coefficient, are used to select informative targets out of various proteins since some studies have observed that the local connectivity of nodes in PPINs plays a crucial role in cellular functions [12,13]. Although these methods give insightful information and achieve successful results, the topology-based analysis of PPINs is far more likely to be heuristic, and this sort of methods cannot identify markers of particular purposes because of ignoring the dynamics of PPINs. Some work relies on gene expression data to predict the informative genes or proteins [14]. However, statistical analysis of gene expression is unable to fully capture the systematical dynamic mechanism, considering it is not capable of investigating the dynamic changes of relationships of proteins in successive PPINs. Also, the statistical methods of gene expression analysis, such as *t*-test and SAM [15,16], tend to be rather arbitrary and sometimes miss the genes with medium expression [14]. To address these problems, this paper tries to analyze the dynamics of PPINs using a deep structure and detects the informative proteins that are related to the development of biological progress.

The proteins exhibiting dramatic structural changes in the set of successive networks are defined as informative proteins which serve as a compensation of the definition of "date" hubs. The topological changes of successive PPINs are smooth and the adjacency networks share certain degree of consistency, hence extracting the consistence and reserving the difference, which makes it possible to find the critical proteins that are extremely important for dynamic processes. The consistency of multiple networks can be represented by their shared hidden features which have been studied as a hot topic recently and our proposed method of multi-layer model differs from traditional methods in both the framework and the aimed task. A canonical method of hidden feature extraction is exponential-family random graph models (ERGMs) [17] which recognizes the complex dependencies within relational data structures. However, this method is irreversible to reconstruct the original network structures. There are several comparative methods of extracting consistent information from multiple graphs, such as the most straightforward average network and the joint non-negative matrix factorization (JNMF) [18]. NMF tries to decompose the original graph to linear combination of basis vectors, and is usually used in clustering problems and graph partition problems. However, the dynamic networks are not linear and a linear solution can only get moderate results.

To this end, this paper proposes a systematical feature memory (SFM) framework that learns a 2-fold systematical feature model in a multi-layer fashion and embedded with a fine-tune procedure that minimizes the reconstruction errors of the whole model. Using the parameters of the model at the previous time point, the current network is reconstructed to filter out the stable structures and the residuals between the adjacent reconstructed networks are analyzed by statistical ranking metrics and finally the informative protein lists are identified. We implemented our method on two datasets: yeast cell cycle data and the data of prostate cancer stages. The reconstruction results are compared with other traditional methods and the informative protein list are verified based on two known gene lists that are proved by current studies to be related to the dynamic processes of cell cycle [19] and prostate cancer development [20]. In summary this work contributes in two ways:

(i) An efficient informative node detection method on dynamic successive networks is proposed which firstly learns a systematical feature memory for these dynamic networks where the common features are extracted utilizing a procedure of multi-layer features learning and fine-tune approximation, and secondly reconstruction analysis is performed which reveals the difference of adjacent networks to locate the informative proteins with most violated structural changes.

(ii) Experimental results show that our strategy of dynamic network construction is superior to the other baseline methods and the SFM is able to reconstruct the dynamic networks with the lowest root mean square error (RMSE) in comparison with other methods since it extracts the consistent hierarchical structures while others do not have any deep insight of the networks.

The rest of the paper is organized as follows: the proposed SFM method and the anomaly detection strategy are described in Section 1; in Section 2 the proposed method is evaluated on two representative datasets and the performance is evaluated based on known knowledge and the number of multiple layers in the model is discussed; finally, the conclusion of this work is given in Section 3.

1 Methods

The framework of this present work is shown in Figure 1. It is assumed that a small part of proteins in the network are associated with the changing of dynamic processes while the majority of the dynamic networks keep stable. Based on this assumption, a method of anomaly detection on dynamic networks is developed to detect the informative proteins.

1.1 Definition of the informative protein detection problem in dynamic networks

PPINs exhibit hierarchical structure and the triggers of structural changes during biological processes can be a small but complex set of molecules [21]. These subgroups can be seen as the hidden factors that affect the topological

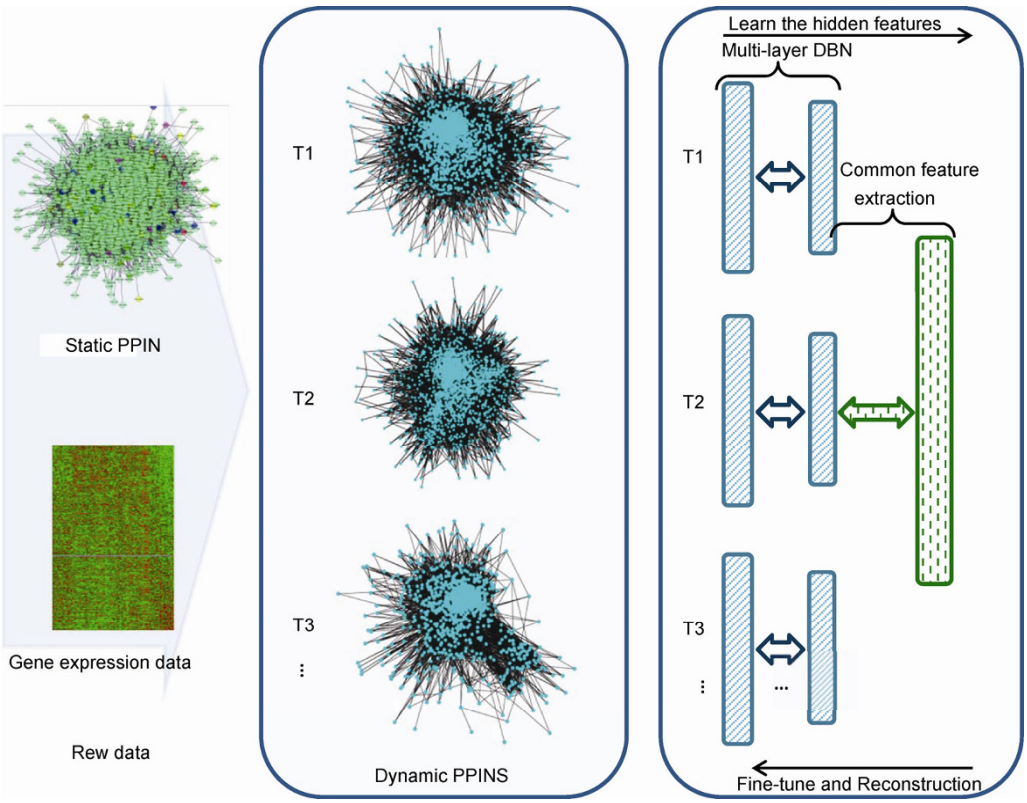

Figure 1 (color online) The framework of this paper. The dynamic networks are constructed based on the static PPIN and gene expression data, and using our proposed SFM model we are able to learn the hidden features of multiple dynamic networks with a fine-tune procedure. Finally with this trained SFM model, the dynamic networks are reconstructed and the reconstruction errors are analyzed to identify the informative proteins.

features of PPINs over time. Given a set of PPINs $\{A_1, A_2, ..., A_T\}$ under T time points which are naturally evolving along the biological processes, the structure of networks is represented by a high-order adjacency matrix in this paper, which can be considered as the reachability of one node to the other in certain steps of random walk. At each time point, the proteins in the PPIN are taken as nodes and each row of the high-order adjacency matrix that a node corresponds to is seen as its feature at that time point, therefore there are T sources about the N nodes. The idea is to capture the interdependent structure between successive networks since adjacent networks share a smooth evolution between them. To rank the most informative proteins, the intuition is that a node will receive low reconstruction error score if its topological structures of neighborhoods are consistent compared to its previous state. So the systematical feature memory method is composed of two parts as illustrated in Figure 2, that is, the deep feature memory (DME) and the associative memory (AME). Once the SFM model is learned, the parameters of DME can be used to reconstruct the dynamic networks. But here we reconstruct the adjacent network using the parameters of the previous time point and it is expected that the stable nodes will receive quite accurate reconstruction while the ones that have unstable topological structures will not be recovered very well because of the mismatch between the parameters of the DME and the input data. These are the most valuable information in anomaly detection problems. Our goal is to find the top-K proteins with greatest reconstruction errors. And the top-K selected informative proteins are validated by candidate protein lists from public databases and literatures.

1.2 Deep feature memory

The DME is a collective model composed of multiple deep belief networks (DBN). To explain the framework of DBN, we should first go through the concept of restricted Boltzmann machines (RBMs), which are stacked one on top of each other to compose the DBNs [22]. RBM is defined as a network of symmetrically-coupled binary random variables or units. As shown in Figure 3, these units can be divided into two groups: the visible variables, $v \in \{0,1\}^{|v|}$, and the hidden variables $h \in \{0,1\}^{|h|}$ ($|\cdot|$ gets the dimension of the object inside it). The visible variables can be the original input or the transformed results from last layer according to the position of current RBM in the whole DBN model. The hidden variables imply the dependencies among the visible variables through their mutual interactional relationships as mimicked by the weighted matrix of W. In RBM, the inter-

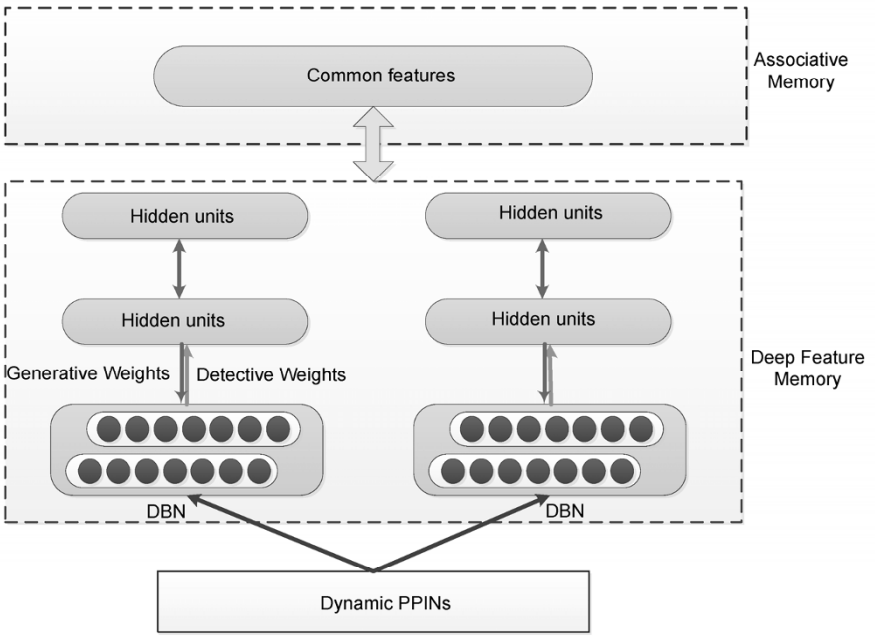

Figure 2 The flowchart of proposed method. There are mainly two parts in this model, i.e., the DME and the AME. The DME is a multiple-layer feature extraction model that learns the features of each PPIN and the AME is another layer that learns the associative features hidden in all of the PPINs.

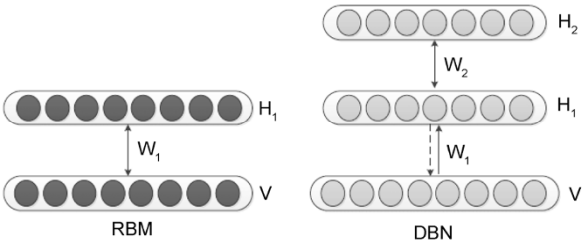

Figure 3 RBM and DBN model.

actions among visible-to-visible variables and among hidden-to-hidden ones are ignored [23]. Hence we get a bipartite graph with completed connections.

The RBM defines an energy function between the visible and hidden layer variables:

$$E(v,h) = -h^{\mathrm{T}}Wv - d^{\mathrm{T}}h - b^{\mathrm{T}}v, \tag{1}$$

where h and v are row vectors in H and V, respectively, b and d are the bias to the visible layer and hidden layer, and W is the weights between two layers. In RBM the purpose of training is to learn the weights and biases between adjacent layers so that the energy function achieves its lowest level. The joint probability distribution of RBM with a normalization factor Z is

$$P(v,h) = \frac{1}{Z}\exp(-E(v,h)). \tag{2}$$

With the restricted conditions, the hidden variables are independent given the visible variables and this property factorizes the individual activation probabilities of a hidden variable as follows:

$$P(h_j = 1 \mid v) = sigmoid\left(d_j + \sum_i W_{ij}v_i\right). \tag{3}$$

Likewise, we have the individual activation probabilities of a visible variable as

$$P(v_i = 1 \mid h) = sigmoid\left(b_i + \sum_j W_{ij}h_j\right), \tag{4}$$

where the *sigmoid* represents the logistic sigmoid function.

To train the probabilistic models, we typically adapt and find the best parameters that maximize the likelihood of the training data. The most straightforward way is to maximize the likelihood following the log-likelihood gradient. However, in the gradient of the log-likelihood, there are terms that are intractable, i.e., the ones that compute the expectations over the joints of variables v and h. There are several ways of dealing with this problem, such as the contrastive divergence (CD) [24] which uses a very short Gibbs chain to estimate the expectation of the joints of v and h. The reliability of CD has been proven by different groups of researchers [25–27].

The RBM model extracts the latent variables hidden in the training data and several RBMs are stacked one on top of others, using the hidden variables derived from lower models as the input, to get deeper layer variables that explain the hierarchical factorizations of PPINs. Given l layers of RBMs as shown in Figure 3B, the joint distribution is

$$P(v, h_1, h_2, ..., h_l) = P(v \mid h_1)P(h_1 \mid h_2)...P(h_{l-2} \mid h_{l-1})P(h_{l-1} \mid h_l). \quad (5)$$

As the variables inside each layer are independent and considering the biases for each layer, we get

$$P(h_{li} = 1 \mid h_{l+1}) = sigmoid\left(b_{li} + \sum_j W_{lij} h_{(l+1)j}\right), \quad (6)$$

$$P(h_{(l+1)i} = 1 \mid h_l) = sigmoid\left(d_{li} + \sum_j W_{lij} h_{(l)j}\right). \quad (7)$$

Running DBNs on each network at each time point, an unsupervised greedy layer-wise feature extraction is pre-trained and the parameter memory is gained. Specifically, each node acts as a sample for the DBN and the goal of learning is to minimize the overall energy of each DBN so that the data distribution can be better captured.

1.3 Associative memory

The hidden structures of each network are learned through the DME phase and the common features of these hidden structures are derived using the associative memory model. Since it is well accepted that the major structures of successive biological networks are stable in biological processes, the unstable topology in networks will be highlighted by the reconstructing procedure following this two-fold model. As shown in Figure 2, the AME phase is interdependent with the DME phase. The input of AME is the pre-learned representations of each protein from each DBN. The pre-learned distributions are sent to the AME which is another separate RBM that is different from the previous ones in DME because the objective of learning is to model the distributions of each protein at different time points. Thus, the distribution of each protein at each time point is a sample and the sample number is decided by the number of time points.

Another main function of the AME phase is to provide the consistent information of each protein as back-propagation for each DME. Using the back-propagation distribution of each protein as the top layer of each DBN, the DME model is fine-tuned to get more coherent distributed description of the original networks. The pseudo code in Algorithm 1 shows how to train the SFM model. In Algorithm 1, lines from 2nd to 6th learn the DME phase, lines from 7th to 10th are the AME part and lines from 11th to 24th are the fine-tune procedure which use the back-propagation information from AME to refine DME. The RBM function in line 4 and line 9 can refer to Bengio's work [28]. Finally when the learning is completed, the parameters of SFM are reserved for further reconstruction analysis.

Algorithm 1 Systematical feature memory (SFM)

Input: 2nd order of adjacency matrices of dynamic networks $A^{(1)},..., A^{(T)}$ and learning rate ε;

Output: Weight matrices $W_l^{(1)},...,W_l^{(T)}$, d_l and b_l

1: Initialize Weight matrices $W_l^{(1)},...,W_l^{(T)}$, d_l and b_l
2: **for all** t in [1:T] (each time point) do
3: **for all** l in [1:L] (each layer of DBN) do
4: $[W_l^t, d_l^t, b_l^t] \leftarrow RBM(h_{l-1}, \epsilon)$;
5: **end for**
6: **end for**
7: compute $P(h_L^t \mid h_{L-1}^t)$ using Eq. 7;
8: **for all** protein i in A do
9: $[W_{top}, d_{top}, b_{top}] \leftarrow RBM(P_i(h_L^t \mid h_{L-1}^t), \epsilon)$;
10: **end for**
11: compute $P(h_L \mid h_{top})$ using Eq. 6;
12: assign $P(h_{L-1}^t \mid h_L^t) = P(h_L \mid h_{top})$;
13: **for all** t in [1:T] (each time point) do
14: **repeat**
15: **for all** l in [1:L] (each layer of DBN) do
16: compute $P(h_{l-1}^{t'} \mid h_l^t)$ using Eq. 6;
17: sample $h_{l-1}^{t'}$ from $P(h_{l-1}^{t'} \mid h_l^t)$;
18: compute $P(h_l^t \mid h_{l-1}^t)$ using Eq. 6;
19: **end for**
20: $W_l^t \leftarrow W_l^t + \epsilon(h_l^t h_{l-1}^{t'} - P(h_l^t \mid h_{l-1}^t))$;
$b_l^t \leftarrow b_l^t + \epsilon(h_l^t - P(h_l^t \mid h_{l-1}^t))$;
$d_l^{(t)} \leftarrow d_l^{(t)} + \epsilon(h_l^t - h_l^{t'})$;
21: **until** the parameters are converged
22: **end for**

1.4 Informative protein detection

As we model the networks with SFM, networks are pulled down to interdependent representations in the two-fold model. The AME phase is able to capture the consistent information of dynamic networks and transfer them into the DME parameters. If the networks are stable in topology, reconstruction using parameters from the previous network will gain the same effect with current parameters; in contrast, it will yield obvious errors if the proteins show great changes in the neighborhoods. We quantify the reconstruction error for each protein i using RMSE which is denoted by Er:

$$Er_i^{(t)} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(A_{ij}^{(t-1)'} - A_{ij}^{(t)'})^2}, \quad (8)$$

where $A_{ij}^{(t-1)'}$ denotes the reconstructed network at time t but using the SFM parameters from time (t–1), $A_{ij}^{(t)'}$ is the reconstructed network at time t using the parameters from time t, and $Er^{(t)}$ is a vector representing the RMSE of proteins at time point t. The $Er^{(t)}$ is ranked to reveal the informative candidates that are more likely to have varied

structures at time point *t* and are expected to play important roles during biological processes.

1.5 Dynamic network construction

It is unlikely to get the dynamic PPINs directly from biological experiments and the usual way is to construct them from static PPI data and time-course information such as the gene expression data. In this paper, the dynamic PPINs are constructed using the method that was proposed in previous publication [29]. In this method, the activity determination of protein and co-related interactions are combined to decide whether two proteins are connected at time *t*.

To decide if a protein is active, a threshold is set for the expression of each gene that is collected under continuous conditions. The active score is defined as

$$AcScore(p) = thr_1(p) \times F(p) + thr_2 \times (1 - F(p)), \quad (9)$$

where $thr_1(p)$ is the mean of the gene expression of protein *p*, which is also denoted as $\mu(p)$, $thr_2(p) = \mu(p) \times \sigma(p)$, where $\sigma(p)$ is the standard deviation of the gene expression of protein *p*, and $F(p) = 1/(1+\sigma(p))$. As seen from eq. (1), *F(p)* is a weight function of $\sigma(p)$ and occurs in the range of (0, 1). An empirical parameter α was set for maintaining the active score *AcScore* within the range of $(\mu(p), \mu(p) + \alpha(\sigma(p)^3)/1+\sigma(p)^2))$. The performance of different empirical α will be discussed in the experimental section.

By setting such an active score threshold, the activity PPI networks Act_t were built for each timestamp:

$$Act_t = \delta_t \delta_t^{\mathrm{T}}, \quad (10)$$

where δ_t is a column vector representing the activity of proteins at time *t* and δ_t^{T} is the transpose of the column vector. Each element in δ_t is determined by the binary threshold function as shown below:

$$\delta_t(p) = \begin{cases} 1, & \text{if } g_t(p) \geqslant AcScore(p), \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

We have used the Pearson correlation coefficient [30] (normalized to the range of 0–1) to calculate the co-expression correlation and build co-regulation protein networks. Since the computation of correlation coefficient requires that expression data be always generated across a period of time, a time window on the original time course expression dataset was set which covered time points previous to and following the current time point *t*. The correlation coefficient matrix at time *t* is denoted as CoE_t. Combining the static PPIN and the activity PPIN provides the dynamic co-regulation protein network at each time point:

$$A_t = CoE_t \circ Act_t \circ Ppi, \quad (12)$$

where *Ppi* denotes the static PPI network adjacency matrix, and ∘ represents element-wise multiplication.

Given the adjacency matrices of networks, the structural difference of networks can be studied in many different ways. The most important point is to incorporate the changes induced by neighbors' behaviors. Hence, we use higher order of the adjacency matrices to mimic random walks on these networks while keeping the non-negative property at the same time.

2 Experiments and results

2.1 Datasets and experiment settings

We used two time-course datasets to evaluate the SFM based informative protein detection method. The first dataset is gene expression data of yeast cell cycle from GSE3431 [31] which is used to construct time course PPINs. GSE3431 is an expression profiling of yeast over three successive metabolic cycles. The overall design of this expression experiment is 12 time intervals per cycle, and approximately 25 min per time interval. Thus, for each gene there are 12 expression values at 12 time points in each cycle. To calculate the instant co-expression correlation coefficient, we choose *t*–1, *t* and *t*+1, as three time points in a snapshot and at each time point there are three successive expression values serving as replicate samples. Further, we adopt another reference cell cycle gene expression data for yeast indexed as GSE7645 to alleviate the bias of expression in the calculation of mean and variance for each gene. In the experiment generating GSE7645, *S. cerevisiae* was cultured under oxidative stress induced by cumene hydroperoxide (CHP) and the transcriptional profile is collected at *t*=0 (immediately before adding CHP) and at 3, 6, 12, 20, 40, 70 and 120 min after adding the oxidant. The static PPIN of yeast was collected from BioGRID dataset for yeast and the cell cycle regulated protein dataset was downloaded from http://www.cyclebase.org/ which will serve as the golden data in validation. We constructed the cell cycle related static PPIN based on these proteins and their first neighbors in BioGRID PPIN. Finally we get a static PPIN with 2069 proteins and 43462 interactions between them.

Another dataset is gene expression data for different stages of prostate cancer from Tomline et al. [20]. Date indexed as GSE6099 examined gene expression profiling of prostate cancer progression from benign epithelium (Benign) to prostatic intraepithelial neoplasia (PIN), to prostate cancer (PCA), and to hormone-refractory metastatic prostate cancer (Met). The co-expression correlations of proteins at certain time are computed by gene expression profiles across three time points including the current time, the previous and the latter time points. So we get four successive networks corresponding to four prostate cancer stages. And the golden data are the representative genes that were found in their work which include 92 genes mainly covering eight

biological functions and have different performance at different stages. The static PPIN of human was collected from Human Protein Reference Database (HPRD), and the static human PPIN of the prostate cancer proteins and their first neighbors in HPRD is constructed and the network contains 2082 proteins and 23098 interactions.

We adopt the CYC2008 human-curated complex dataset as benchmark data [32] to evaluate the accuracy of our constructed dynamic PPINs. CYC2008 is a comprehensive catalogue of manually curated 408 heteromeric protein complexes in *S. cerevisiae* reliably backed by small-scale experiments from literature.

Since we cannot get the precise time from the known knowledge when each protein comes into effect to change the biological progression, we take all the informative proteins found at each network into an all-covered unique list and compare with the golden list to validate the performance of the SFM model. The precision of the model is defined as $Prec = TP/(TP+FP)$, where *TP* (True Positive) is the number of the predicted informative proteins matched with known proteins in golden datasets, and *FP* (False Positive) is the number of the unmatched proteins that are found in the predicted list.

2.2 Evaluation of SFM

2.2.1 Comparison of different dynamic networks reconstruction methods

In the SFM model, the visible input variables were chosen as the high order of the adjacency matrices, in this case the 2nd-order was used, and the self-transmissions were ignored which meant the diagonal elements in each matrix were set to 0. We built the separate training DME as a 4-layer model and the parameter choosing will be discussed in the following subsection. We compared our method with two basic reconstruction methods and also with the original DBN to verify the effectiveness of our method.

The baseline methods include Joint NMF (JNMF) method, the straightforward average network and the original DBN method. The JNMF method learns a common base matrix from different sources that best approximates the original sources. It is often used in clustering problems and dimension reduction problems. In our experiments the prior low dimension of JNMF was set as 500 by which the approximation to the original data generally achieved the best position. And the method which adopts the average network, denoted as AVG in the following content, simply extracted the average of the 2nd order adjacencies of the series of dynamic networks. Compared with our SFM, the DBN method just processes our dynamic networks through one straightforward deep structure of three layers to get the deep representations and derive the reconstruction errors using the same parameters on different networks. By comparing the RMSEs, it is easy to see in Figure 4 that the SFM method obtains the best reconstruction while the AVG gets the worst in all of the four methods. As discussed in Section 1, the SFM model is to extract interdependent representations for the smoothly evolving dynamic networks. In the SFM model the networks are first trained in separate deep structures to get their representative deep feature models and then the common features are processed in the AME to get a mutual-restrained representation based on which the DME is further tuned. The JNMF is analogous to one layer feature extraction model that does not fit in to its best within this scenario. In addition, our method surpasses the traditional DBN which considers all the networks identically and shows the promising results of the systematical deep structure.

2.2.2 Precision of the informative protein detection

To get the informative proteins, we ranked the RSMEs of proteins at each time point separately and then put the identified targets into the final list that would only keep the unique proteins. Since it is not possible currently for the collected golden data sets to reveal the specific time when each protein shows its critical effect on the biological progress, we take every protein equally to be informative in the

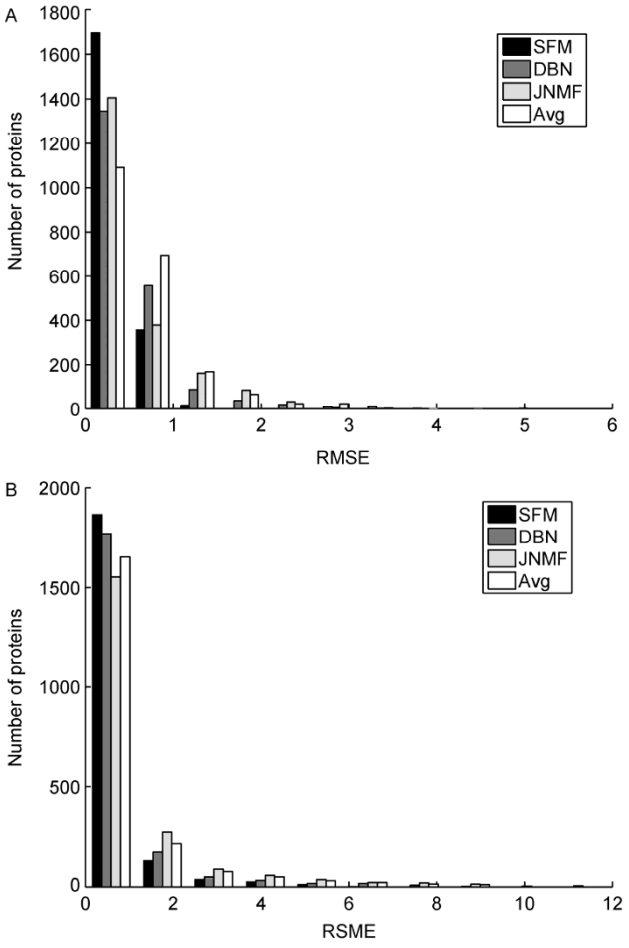

Figure 4 Comparison of RSME. A, Yeast cell cycle dataset. B, Prostate cancer dataset.

evaluation part as if it shows up at one time point. We varied the top k that was chosen from each ranked list at each time point, getting the precisions as shown in Table 1. We see that our informative protein detection method based on SFM model is far more precise than the other three compared methods. We also recognize that with the number of k increasing the precision is falling gradually rather than dramatically, since not only the number of identified proteins is increasing but also the matched proteins with the golden lists.

2.2.3 *Parameter settings*

Initially, we analyzed the effect of change in parameters on our dynamic network construction method. We performed spectral method to detect dynamic functional modules at 12 time points and compared the results with the CYC2008 dataset. The *Prec*s of the results under different parameter settings have been compared as shown in Table 2. From the results of comparison, it was obvious that with α fixed at 1.5, the precision of the functional module detection achieved the highest score. Thus, in this work, this prime parameter setting has been used.

In the SFM model, one of the most important superiorities is the deep feature representation. Here we want to analyze how deep of a DME model should be to achieve the best performance as a whole model. We tried different numbers of layers for the DME phase and run 10 times for each situation to get the average precisions. The precisions when top 10 proteins are chosen from the ranking results are shown in Figure 5. When we increased the number of layers

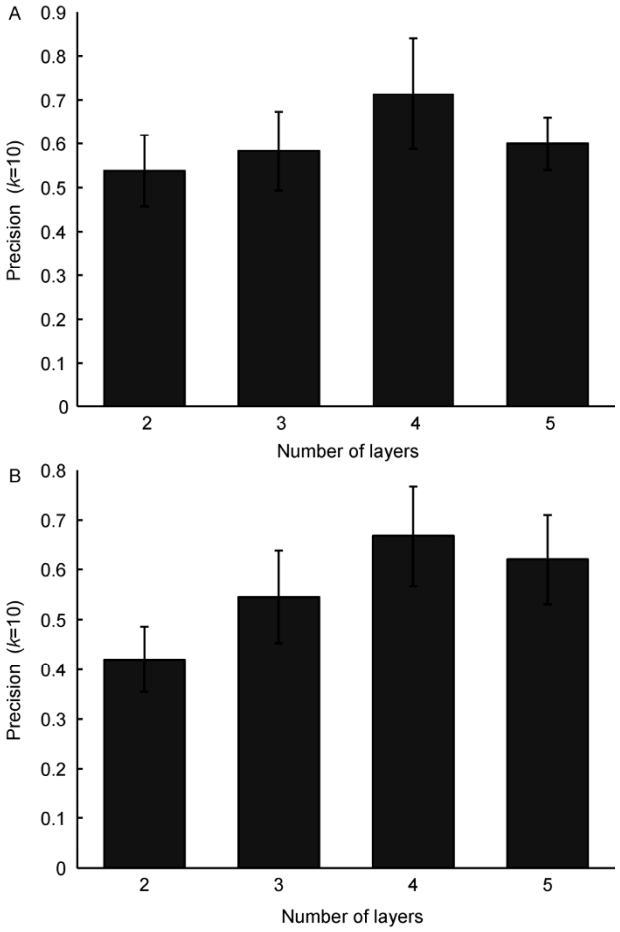

Figure 5 Performance of SFM w.r.t number of layers.

Table 1 Precision of different methods

k		5	10	15	20	25	30	35	40	45	50
Yeast	SFM	0.7	0.714	0.643	0.649	0.628	0.621	0.582	0.537	0.528	0.526
	DBN	0.538	0.52	0.531	0.488	0.451	0.439	0.427	0.425	0.389	0.402
	JNMF	0.417	0.429	0.412	0.39	0.327	0.302	0.28	0.265	0.275	0.248
	Avg	0.286	0.217	0.194	0.233	0.216	0.169	0.169	0.165	0.161	0.14
Prostate	SFM	0.6	0.667	0.621	0.611	0.583	0.55	0.536	0.471	0.433	0.408
	DBN	0.583	0.526	0.5	0.487	0.46	0.429	0.408	0.4	0.402	0.393
	JNMF	0.385	0.381	0.313	0.317	0.288	0.295	0.291	0.272	0.23	0.219
	Avg	0.25	0.238	0.152	0.135	0.118	0.129	0.127	0.101	0.102	0.114

Table 2 Parameter settings of dynamic networks construction, $\alpha \in [0.5-3.5]$

α	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	Average
0.5	0.27	0.35	0.325	0.325	0.375	0.35	0.2	0.375	0.225	0.4	0.275	0.325	0.316
1	0.34	0.483	0.483	0.317	0.35	0.417	0.283	0.45	0.4	0.333	0.333	0.317	0.376
1.5	0.533	0.517	0.483	0.4	0.417	0.417	0.4	0.55	0.583	0.533	0.567	0.533	0.494
2	0.33	0.4	0.48	0.35	0.32	0.32	0.38	0.46	0.49	0.49	0.46	0.5	0.415
2.5	0.47	0.387	0.445	0.328	0.345	0.312	0.32	0.478	0.495	0.545	0.478	0.47	0.423
3	0.423	0.394	0.437	0.28	0.287	0.316	0.316	0.451	0.473	0.501	0.423	0.416	0.393
3.5	0.39	0.371	0.421	0.315	0.29	0.265	0.303	0.421	0.528	0.453	0.415	0.415	0.38

from 2 to 4, the performance got better accordingly, but when it was raised to 5, the precision fell, on the contrary. We thus assert that it is not true that the more layers the better performance a model gets. Meanwhile, more layers mean more computation complexity; hence we chose 4-layer DME to learn the SFM.

3 Conclusion

In this paper, a systematic deep feature model was proposed to study the structural variability of successive dynamic PPINs. In the SFM model, the respective deep feature of dynamic networks and also the interdependent relationships were modeled by two sub-modules, i.e., DME and AME. With these models that were learned by SFM, the original dynamic networks were reconstructed using the parameters from its previous network and by comparing the two reconstructed networks, the informative proteins were identified. We evaluated our work on two representative datasets, the yeast cell cycle and the human prostate cancer stages datasets. By comparing the reconstruction performance with other traditional methods, we saw that the SFM can best recover the dynamic networks. Besides, the ranking results of informative proteins from SFM were compared with results from JNMF reconstruction method and the comparison of results showed that SFM identified more proteins of critical value to the biological processes which can provide valuable information for further study such as medicine design, clinical diagnosis and disease treatments.

One thing worth mentioning is that, as we compare the RSME of yeast cell cycle and prostate cancer in Figure 4, we see that SFMs show better performance on the yeast cell cycle dataset. The reason might be that the sparsity of networks could influence the reconstruction ability of SFM as we know that the human PPIN is more isolated than yeast PPIN. That is one of the problems we currently try to solve. Moreover, the fact that a few proteins among the unmatched protein list are truly relevant to the biological process inspires an interesting idea that the system analysis of dynamic networks should be done to reveal groups of critical proteins with the same or relative functional roles in the dynamic mechanism. In the future, we will focus more on a system level study of the dynamic networks.

The authors declare that they have no conflict of interest regarding the publication of this article.

This work was supported by National Natural Science Foundation of China (30970780), Ph.D. Programs Foundation of Ministry of Education of China (20091103110005), the Project for the Innovation Team of Beijing, National Natural Science Foundation of China (81370038), the Beijing Natural Science Foundation (7142012), the Science and Technology Project of Beijing Municipal Education Commission (km201410005003), the Rixin Fund of Beijing University of Technology (2013-RX-L04), and the Basic Research Fund of Beijing University of Technology.

1 Overby CL, Tarczy-Hornoch P. Personalized medicine: challenges and opportunities for translational bioinformatics. Pers Med, 2013, 10: 453–462
2 Olson S, Beachy SH, Giammaria CF, Berger AC. Integrating Large-Scale Genomic Information Into Clinical Practice: Workshop Summary. Washington, DC: National Academies Press, 2012
3 Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. New Biotechnol, 2012, 29: 613–624
4 Chang X, Xu T, Li Y, Wang K. Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date'' and 'party' hubs. Sci Rep-Uk, 2013, 3: 1691
5 Komurov K, White M. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. Mol Syst Biol, 2007, 3: 110
6 de Lichtenberg U, Jensen LJ, Brunak S, Bork P. Dynamic complex formation during the yeast cell cycle. Science, 2005, 307: 724–727
7 Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 2004, 430: 88–93
8 Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotech, 2009, 27: 199–204
9 Wang H, Li M, Wang J, Pan Y. A New Method for Identifying Essential Proteins Based on Edge Clustering Coefficient. In: Chen J, Wang J, Zelikovsky A, eds. Bioinformatics Research and Applications. Berlin Heidelberg: Springer, 2011. 87–98
10 Wang J, Peng W, Wu FX. Computational approaches to predicting essential proteins: a survey. Proteomics Clin Appl, 2013, 7: 181–192
11 He X, Zhang J. Why do hubs tend to be essential in protein networks? Plos Genet, 2006, 2: e88
12 Barabasi A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet, 2004, 5: 101–113
13 Wang Z, Lucas FA, Qiu P, Liu Y. Improving the sensitivity of sample clustering by leveraging gene co-expression networks in variable selection. BMC Bioinformatics, 2014, 15: 153
14 Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn, 2002, 46: 389–422
15 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response (vol 98, pg 5116, 2001). Proc Natl Acad Sci USA, 2001, 98: 10515–10534
16 Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics, 2002, 18: 546–554
17 Handcock MS, Robins G, Snijders TA, Moody J, Besag J. Assessing degeneracy in statistical models of social networks. CSSS working paper, 2003: 39
18 Ge L, Gao J, Yu X, Fan W, Zhang A. Estimating local information trustworthiness via multi-source joint matrix factorization. In: 12th IEEE International Conference on Data Mining, 2012. 876–881
19 The cyclebase database. http://www.cyclebase.org/
20 Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, Shah RB, Chinnaiyan AM. Integrative molecular concept modeling of prostate cancer progression. Nat Genet, 2007, 39: 41–51
21 Arnau V, Mars S, Marin I. Iterative cluster analysis of protein interaction data. Bioinformatics, 2005, 21: 364–378
22 Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507
23 Rumelhart DE, McClelland JL. Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations. Cambridge, MA, USA, 1986
24 Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput, 2006, 18: 1527–1554
25 Yuille A. The convergence of contrastive divergences. Advances in

Neural Information Processing Systems 17. Cambridge, MA, USA, 2005. 1593–1600
26 Bengio Y, Delalleau O. Justifying and generalizing contrastive divergence. Neural Comput, 2009, 21: 1601–1621
27 Sutskever I, Tieleman T. On the convergence properties of contrastive divergence. In: International Conference on Artificial Intelligence and Statistics, 2010. 789–795
28 Bengio Y. Learning deep architectures for AI. Found Trends Mach Learn, 2009, 2: 1–127
29 Zhang Y, Du N, Li K, Jia K, Zhang A. Co-regulated protein functional modules with varying activities in dynamic PPI networks. Tsinghua Sci Tech, 2013, 18: 530–540
30 Bhardwaj N, Lu H. Correlation between gene expression profiles and protein-protein interactions within and across genomes. Bioinformatics, 2005, 21: 2730–2738
31 Tu BP, Kudlicki A, Rowicka M, McKnight SL. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. Science, 2005, 310: 1152–1158
32 Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res, 2009, 37: 825–831